DOCUMENT RESUME

ED 106 309 TM 004 465

AUTHOR Kaskowitz, David; Stallings, Jane
TITLE An Assessment of Program Implementation in Project Follow Through.
INSTITUTION Stanford Research Inst., Menlo Park, Calif.
PUB DATE [Apr 75]
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975) For related documents, see TM 004 572, 573 and ED 085 100; Tables 1 and 2 may reproduce poorly

EDRS PRICE MF-$0.76 HC-$1.95 PLUS POSTAGE
DESCRIPTORS Academic Achievement; *Classroom Observation Techniques; Classroom Research; Comparative Analysis; Data Analysis; Economic Disadvantagement; Educational Innovation; *Elementary Education; Evaluation; Federal Programs; *Intervention; Models; *Program Effectiveness; Test Reliability
IDENTIFIERS Classroom Observation Instrument; Implementation; *Project Follow Through

ABSTRACT

Methodological issues and results described in this paper originated from a Stanford Research Institute (SRI) evaluation of classroom observation data collected in Spring 1973. The main question addressed in this evaluation was whether each of seven Follow Through sponsors had successfully implemented his program in a variety of sites. The steps in the evaluation of implementation included: (1) a determination of the essential program components, (2) a translation of these components in terms of observable phenomena, (3) a measure of the phenomena, and (4) a standard by which to judge implementation. The SRI approach was formulated in the context of the SRI Classroom Observation Instrument in liaison with the Follow Through sponsors. Only program components that could be translated into observation variables were included in the analysis. To focus on the innovative aspects of the Follow Through programs, the standard established to assess implementation was based on comparisons with Non-Follow Through Classrooms. Implementation scores were derived for individual observation variables, and a total implementation score was derived for each classroom. While all sponsors' classrooms had mean total implementation scores that were significantly different from the Non-Follow Through scores, there were differences in the pattern of implementation among sponsors. (Author/RC)

AN ASSESSMENT OF PROGRAM IMPLEMENTATION IN PROJECT FOLLOW THROUGH

By: David Kaskowitz
Jane Stallings

Presented to the

American Educational Research Association
1975 Annual Meeting,
Washington, D.C.,
March 31-April 3, 1975

AN ASSESSMENT OF PROGRAM IMPLEMENTATION IN PROJECT FOLLOW THROUGH

Introduction

The methodological issues and results described in this paper originated from an SRI evaluation of classroom observation data collected in Spring 1973 (Stallings and Kaskowitz, 1974). The main question addressed in this evaluation was whether each of seven Follow Through sponsors had successfully implemented his program in a variety of sites.

Project Follow Through was established by Congress in 1967 (the legislative authority for Project Follow Through was the Economic Opportunity Act of 1964, as amended) when it became apparent that a program was needed in the early grades of public school that was compatible with Project Head Start goals and approaches and, therefore, would provide a comparable educational program for economically disadvantaged children over a longer period of time.

Follow Through was originally set up in a "planned variation" research design; that is, the goal was to examine the differential effectiveness of programs based on divergent educational and developmental theories. The program began when 22 educational researchers, later called Follow Through program sponsors, were invited by the government to submit plans for establishing their various programs in public schools to test the potential of these programs for improving the educational achievement of economically disadvantaged children. Eleven of the sponsors had developed and tried their educational concepts in university settings, eight were affiliated with private research institutes, and three were community development programs.

A major issue in the evaluation of any innovative educational program is whether the program can be successfully implemented under the variety of conditions encountered in the field. For policy purposes, disseminators and potential recipients of innovative programs need to know whether the educational components of a model can be implemented widely and with consistent success in a reasonable amount of time, for reasonable cost and effort. The issue of implementation is also most important in assessing program effects. Just because a program is labeled Brand X does not guarantee that the essential properties of Brand X are present. Conclusions about the effects of a program must

be based on the assumption that the program is indeed implemented. Although the information that a program is not implemented may be of interest in its own right, an evaluation of the effects of such a "program" is of questionable value. In fact, such an evaluation could be misleading and unfair, since conclusions stated in terms of the particular effects may not really pertain to that program.

The SRI Classroom Observation Instrument was used to gather data about classroom environment and processes. The instrument was initially developed in 1969 with the assistance of eight Follow Through sponsor representatives with a goal of being flexible enough to record the significant features of a variety of program components. The instrument consists of five sections:

- Classroom Summary Information (CSI)--The CSI is filled out once each day. It identifies the sponsor and teacher and provides information on the number of teachers, aides, volunteers, and students, and the class duration.
- Physical Environment Information (PEI)--The PEI is filled out once each day. It describes the seating patterns and the presence and use of equipment and materials.
- Classroom Check List (CCL)--A CCL is filled out about four times an hour. It provides infomration on the grouping of children and teaching staff and activities in the classroom.
- Preamble (PRE)--A Preamble is filled out after each CCL. It describes the activity and role of the person who is the focus of the Five-Minute Observations.
- Five-Minute Observation (FMO)--The FMO is filled out after each Preamble. It uses coded sentences to describe the interactions occurring in the classroom. The information includes the parties to the interaction, the type of interaction, and the quality of the interaction.

The following section discusses the major methodolcgical issues related to measuring implementation. This is followed by a description of the methods used in the SRI study, a discussion of the reliability of the implementation scores, and an overview of the results.

The Methodological Issues

The analytic approach to the evaluation of implementation was based on both theoretical and practical considerations. The answers to such basic questions as

- How does one define the term "implementation?"
- What program components should be included?
- What sort of implementation measure should be used?

were necessarily shaped within the context of the available resources, including the observation instrument and sponsor input.

In general, a program may be said to be implemented in a particular classroom or site to the extent that essential program components are present. The approach to measuring implementation requires a specification of essential program components, a means of obtaining a measure of the degree to which each component is present, and an explicit statement of the criteria for implementation.

The information coded in the Classroom Observation Instrument determined the scope of the current analysis. Many of the significant features of the Follow Through programs are included, such as type and use of staff, use of materials and equipment, classroom configurations and activities, and frequency and type of interactions. For most Follow Through sponsors, and specifically for those included in the evaluation, the classroom environment is the primary medium in which the program is translated into a child's educational experience. Components not directly observable in the classroom were necessarily excluded from the analysis.

For many Follow Through programs, some of the major educational components cannot be directly measured. For example, a goal of Far West Laboratory is to have its teachers establish environments in which a child can search for solutions to his problems in his own way and can risk, guess, and make discoveries without serious negative psychological consequences. It is not possible to measure directly the extent to which such an environment has been established. Some observable behaviors may be examined, such as the frequency of question asking or the variety of activities occurring, but these are feeble proxies. To the extent that a sponsor's program cannot be expressed in terms of specific and observable components, it is not possible to measure objectively whether the program is implemented.

For certain physical manifestations of program impact, such as the presence of specified materials, the presence of a specified number of teachers and aides, and the presence of a specified arrangement of the classroom, an all-or-none measure is sufficient for assessing implementation. Most of these all-or-none components are necessary, but not sufficient, for program implementation. For example, all the specified equipment may be present, but it may not be used or it may be used incorrectly.

Other implementation components require some measure of frequency or relative frequency. For example, most Follow Through programs call for the teaching staff to work with small groups of children more frequently than with the entire class. Other programs call for the children to work independently of the teaching staff for the most part. Implementation in these cases is related to the frequency or relative frequency of occurrence rather than to the presence or absence of the component.

The criterion for determining the degree of program implementation is self-evident for components that are all-or-none in nature. If specified equipment is present, the requisite number of aides are present, or the room is in the specified arrangement, then the particular component is implemented. For the more interesting and more numerous components that are measured by rates of occurrence or relative frequency, some standard for assessing implementation must be established.

To determine what should be occurring, we need to derive the value of the measure for each component that would be observed in a perfectly implemented classroom. Unfortunately, such specific criteria cannot be deduced from the educational philosophy of even the most structured program. A more realistic approach is to set the standard by observing the program classrooms that are designated as the most exemplary of an implemented classroom. The problem with this approach is that no one class may be fully implemented on all the requisite components. Thus, observations on several classrooms might have to be pieced together to derive a composite standard, or the protocols for several classrooms that are judged to be well implemented could be averaged to derive a standard.

Another approach is to use observations of "conventional" classrooms as a base from which the program classrooms will differ if the requisite program is implemented. Since this approach focuses on the innovative aspects of a program, educational components that do not differentiate a sponsor's classroom from a conventional classroom should not be examined, since the component would probably be evident even if the program were not implemented. If a program is intended to be innovative, the evaluation of implementation should be based only on aspects of the program that are innovative.

The problem with this approach lies in assessing the direction and degree of difference from the conventional classroom necessary for implementation to be achieved. The direction can generally be derived either from examination of a sponsor's educational theory or more directly from a direct inquiry of the sponsor. From the experience of the current evaluation, sponsors can easily determine the direction of differences from the conventional classroom that a well implemented classroom would exhibit.

The question of the degree of difference that is indicative of implementation has not been satisfactorily resolved. No sponsor's model is specific enough to determine such criteria. In the absence of reliance on a theoretical approach to establishing the criteria, one is usually left with establishing criteria based on statistical significance of differences. There may be no alternative to this strategy, but the statistical significance of any differences detected may not correspond to educational significance.

One approach that may resolve this problem to some degree would incorporate both the standard based on observations of well-implemented classrooms and the standard based on observations of conventional classrooms. An implementation scale could then be derived that places the well-implemented standard at one extreme and the conventional standard at the other.

Once an implementation measure has been adopted, its reliability must be examined. The reliability of the implementation measure depends on the reliability of the observation instrument and on the number of classrooms and class days per classroom designated for observation. The accuracy of the implementation measures used in the SRI evaluation is assessed in a later section of this paper.

A final methodological issue that should be mentioned pertains to the aggregation of implementation measures on individual components to derive a total implementation measure. Such a measure should ideally take into account the differential importance of the educational components as well as the dependence among the components. Program sponsors may be able to categorize components into gross categories of relative importance, but they probably cannot make any finer differentation. Assigning an equal weight to each component within these categories and eliminating the less important components from the assessment of implementation appears to be the most satisfactory solution to the aggregation question.

The Sample and Observation Procedure

The sample for Spring 1973 included teacher observations and child observations in the 36 Follow Through projects identified in Table 1. The sample included seven Follow Through project sponsors and five projects within each sponsor except for the University of Arizona, where observations were conducted in six projects. Both first grade and third grade classrooms were included. Observations were conducted for approximately four Follow Through and for one Non-Follow Through (NFT) classroom at each grade level.

Table 1

CLASSROOM OBSERVATION SAMPLE, SPRING 1973

Sponsor and Sites		Number of Follow Through Classes Observed: First Grade	Third Grade
Far West Laboratory for Educational R&D			
0201	Berkeley, Calif.*	4	4
0204	Duluth, Minn.*	4	4
0207	Lebanon, N.H.	4	4
0209	Salt Lake City, Utah	4	4
0213	Tacoma, Wash.	4	4
University of Arizona			
0305	Des Moines, Iowa	4	4
0307	Fort Worth, Texas*	4	4
0308	LaFayette, Ga.	3	4
0309	Lakewood, N.J.	4	4
0311	Newark, N.J.	4	4
0316	Lincoln, Nebraska	4	4
Bank Street College			
0502	Brattleboro, Vermont	3	3
0504	Fall River, Mass.	4	4
0506	New York City, P.S. 243K	4	4
0508	Philadelphia II, Pa.*	4	4
0510	Tuskegee, Ala.*	4	4
University of Oregon			
0703	E. St. Louis, Ill.	4	4
0707	New York City, P.S. 137K	3	3
0708	Racine. Wisc.	4	4
0711	Tupelo, Miss.*	4	4
0719	Providence, R.I.	4	4
University of Kansas			
0801	New York City, P.S. 77X*	2	2
0803	Philadelphia VI, Pa.*	4	4
0804	Portageville, Mo.*	4	3
0806	Kansas City, Mo.	4	4
0807	Louisville, Ky.	4	4
High Scope Educational Research Foundation			
0901	Greenwood, Miss.*	4	4
0902	Ft. Walton Beach, Fla.*	4	4
0903	New York City, P.S. 92M	3	4
0906	Greeley, Colo.	3	3
0907	Denver, Colo.	4	4
Education Development Center			
1101	Burlington, Vermont	4	4
1103	Philadelphia IV, Pa.*	4	4
1106	Paterson, N.J.*	4	4
1107	Rosebud, Texas	3	3
1108	Smithfield, N.C.	4	2
	Total	136	135

*These sites have been observed previously.

Each classroom was observed for three days. Two days were devoted to observing the teacher and other adults in the classroom (adult/activity focus), and the remaining day was devoted to observing individual children in each classroom (child focus). Generally, only the Five-Minute Observation data, which record the interactions in the classroom, are affected by changes in the focus of observation; the interpretation of other variables related to classroom activities and configurations does not depend on the focus of observation.

Two observers were hired locally in each of the 36 projects. Each observer attended a seven-day training session conducted by the SRI training staff. Satisfactory completion of the training program was required before an observer was allowed to conduct observations in the field.

Method of SRI Implementation Evaluation

The first step in assessing classroom implementation was to describe in detail each educational model. These descriptions were reviewed by sponsors and revised according to each sponsor's specifications. The next step was to translate the descriptions into operational terms by creating variables from the codes used on the observation instrument to describe the critical elements of each sponsor's program. Each sponsor was sent a list of critical variables and asked to rate each one in terms of its importance to the program and its expected frequency relative to conventional classrooms. Only variables that were considered critical and that should occur more often than in conventional classrooms were used as implementation variables. (None of the critical variables was expected to occur less frequently than in conventional classrooms, since generally what is critical is expressed in terms of what should be happening. Because of the relationships among the variables, the high frequency of a critical variable will usually mean the low frequency of a variety of other variables.)

Sponsors were also given the opportunity to add observation variables that they considered important to their programs. The variables selected to measure implementation are presented in Table 2 for each model. The variables chosen by the sponsors overlap considerably; however, it is the unique mix of variables that makes the models different from one another. (An analysis of differences among sponsors was carried out in parallel to the implementation study. Results are reported in Chapter VI of the SRI evaluation report: Stallings and Kaskowitz, 1974.)

The third step was to establish a standard for each critical variable by which the degree of implementation could be measured. Both a first grade and a third grade classroom designated by each sponsor as implemented

Table 2

LIST OF CRITICAL VARIABLES SELECTED BY SPONSORS

Variables No.	Description	Far West Labs	University of Arizona	Bank Street	University of Oregon	University of Kansas	High Scope	EDC
24	Child selection of seating and work groups	X	X	X			X	X
25	Games, toys, play equipment present	X	X	X			X	X
39	General equipment, materials present	X				X		X
65	Guessing games, table games, puzzles		X			X	X	
66	Numbers, math, arithmetic	X	X	X	X	X	X	X*
67	Reading, alphabet, language development	X	X	X	X	X	X	X*
70	Sewing, cooking, pounding		X				X	
71	Blocks, trucks						X	
74	Practical skills acquisition	X**		X				
83	Wide variety of activities, over one day	X	X	X			X	X
86	Teacher with one child	X	X	X			X	
87	Teacher with two children			X				
88	Teacher with small group		X	X	X	X	X	
92	Aide with one child	X		X				
94	Aide with small group		X		X	X	X	
114	One child independent	X	X	X				X
115	Two children independent							X
116	Small group of children independent	X		X				X
239	Math or science equipment/Academic Activities	X		X			X	X
240	Texts, workbooks/Academic Activities				X	X		
343	Child to adult, all verbal except response			X				
344	Individual child verbal interactions with adult	X	X	X	X	X	X	X
350	Child questions to adults	X	X	X			X	X
363	Child group response to adult academic commands/requests or direct questions				X			
372	Child presenting information to a group			X			X	
375	Adult instructs an individual child	X						X
376	Adult instructs a group				X			
390	Adult task-related comments to children						X	X
394	All adult acknowledgment to children	X		X	X	X	X	
398	All adult praise to children		X		X	X		
412	Adult feedback to child response to adult academic commands/requests, questions				X	X		
420	Adults attentive to a small group	X			X		X	
421	Adults attentive to individual children	X	X	X		X	X	
423	Positive behavior, adults to children	X	X	X				
435	Total academic verbal interactions				X			
438	Adult communication or attention focus, one child	X		X		X	X	X
440	Adult communication or attention focus, small group				X		X	
444	Adult movement	X						
450	All child open-ended questions							X
451	Adult academic commands/requests and direct questions to children				X	X		
452	Adult open-ended questions to children	X	X	X			X	X
453	Adult response to child's question with a question						X	X
454	Child's extended response to questions		X	X				
456	All child task-related comments	X	X				X	
457	All adult positive corrective feedback	X			X	X		
460	All child positive affect	X	X					X
469	All adult reinforcement with tokens					X		
509	Child self-instruction, academic				X*			
510	Child self-instruction, objects			X			X	X
513	Child task persistence			X		X	X	
514	Two children working together, using concrete objects						X	
515	Small group working together, using concrete objects			X			X	
516	Social interaction among children	X		X				X
574	Child movement	X						X
599	Child self-instruction, nonacademic	X	X				X	
	Total number of Critical Variables	28** 27*	21	27	16** 17*	17	29	20** 22*

*Third grade only.

**First grade only.

had been observed to generate an implementation standard to which the sponsor's classrooms should conform. Approximately 36 Non-Follow Through classrooms for each grade level, one from each site, had also been observed.

After the observations had been made, it was the opinion of SRI professionals that the criterion classroom observations of the "ideally" implemented classroom were not adequate for establishing a standard because, although each teacher was an excellent example of a sponsor's model, no sponsor was willing to guarantee that all other teachers would or should perform exactly like the one selected as criterion. Instead of specifying the ideal implementation standards for each model, SRI used the relatively conventional Non-Follow Through classrooms to set the standard from which a Follow Through classroom should differ if it is implemented.

The technique used to derive implementation scores was a refinement of one used in several past SRI evaluations (Stallings, 1973; Stearns, Preecs, and Steinmetz, 1973). A nonparametric scaling technique was selected over one that required assumptions concerning the distribution of the Non-Follow Through classrooms because of the variety of distributions that were encountered for the NFT classrooms. As an illustration, Figures 1 and 2 display histograms for first grade NFT classrooms on two implementation variables. The distribution in Figure 1 is close to the familiar bell-shaped normal curve, and the distribution in Figure 2 has a reversed J shape with some extreme outliers. Any parametric approach that may appear to be appropriate for one distribution may be entirely misleading for another type. Also, the nonparametric scaling procedure chosen tends to be less sensitive to outliers than a more conventional approach that might depend on the mean and standard deviation.

An implementation standard was obtained by dividing the Non-Follow Through distribution into equal parts on a percentage basis. Each Follow Through classroom could be assigned an implementation score on a particular variable, depending on the position of the Follow Through classroom value relative to the distribution of the Non-Follow Through classrooms. Several alternatives to the number of divisions of the Non-Follow Through distribution were considered, and it was decided that the quintiles of the NFT distribution that divide the distribution into five parts were adequate for deriving implementation scores.

The relative frequency distribution of NFT classrooms was derived for each implementation variable separately for first and third grade. The four quintiles of the distribution--corresponding to the 20th, 40th, 60th, and 80th percentiles--were derived. For example, there were 35

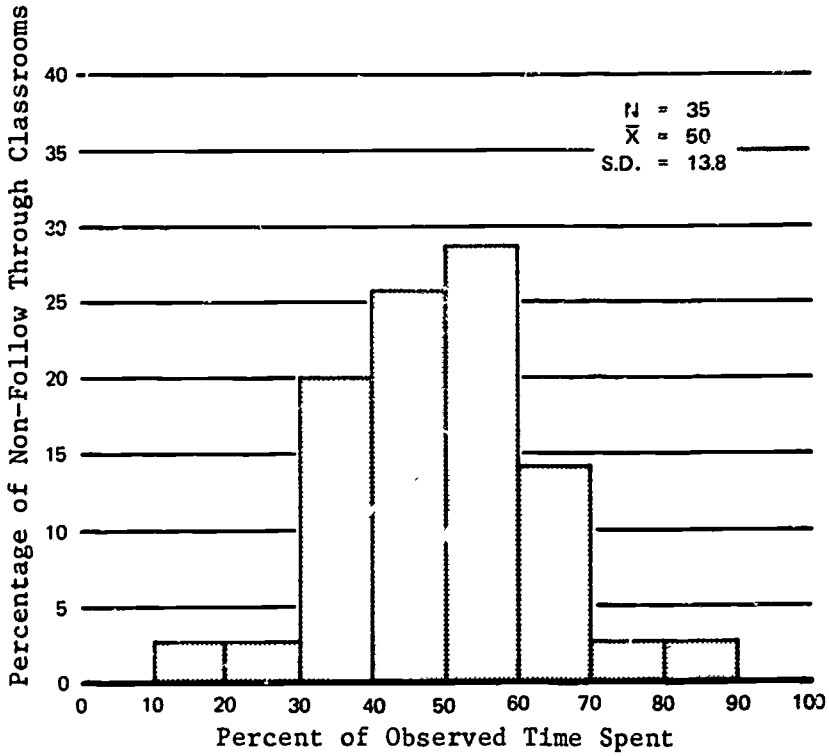

FIGURE 1 HISTOGRAM OF FIRST GRADE NON-FOLLOW THROUGH CLASSROOMS SHOWING PERCENT OF CHILD TIME SPENT IN READING, LANGUAGE (VAR. 67)

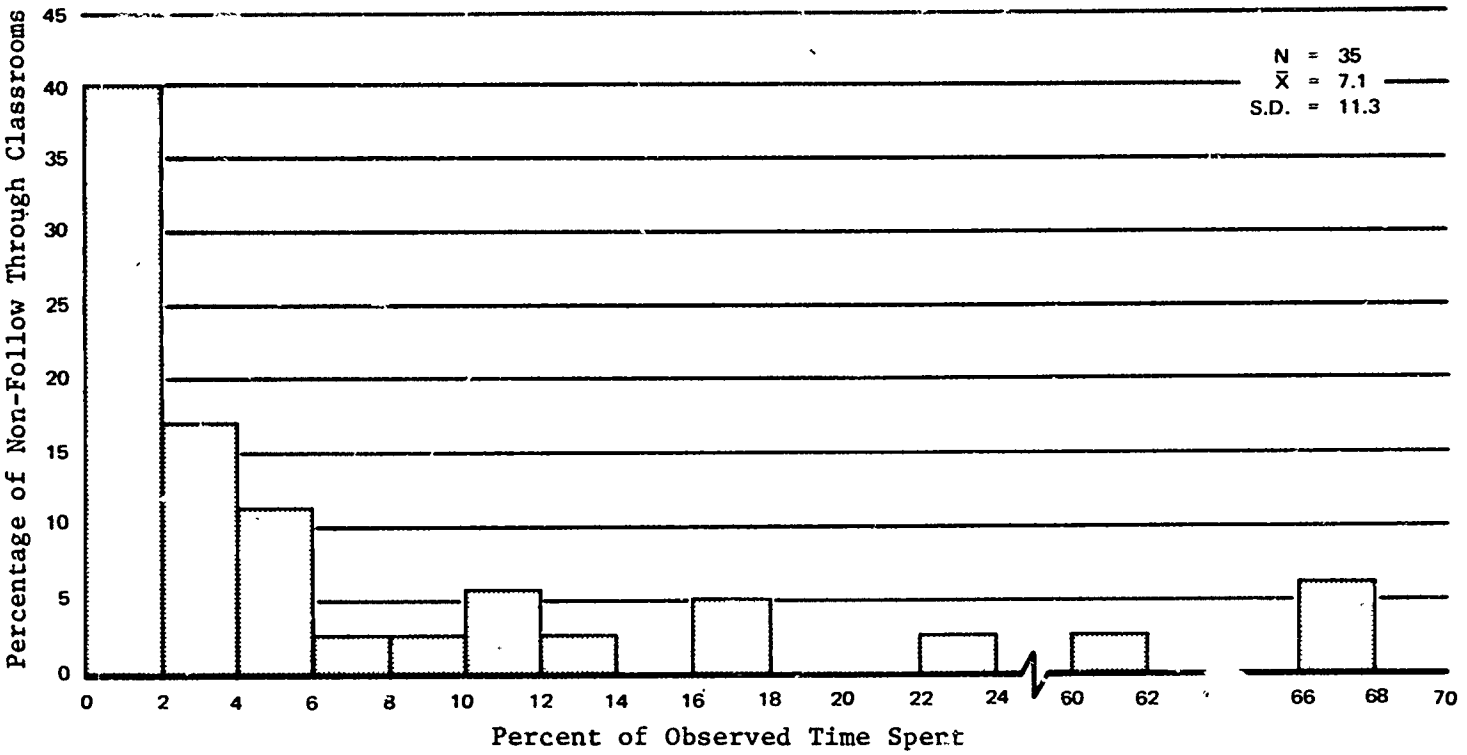

FIGURE 2 HISTOGRAM OF FIRST GRADE NON-FOLLOW THROUGH CLASSROOMS SHOWING PERCENT OF TIME TEACHER SPENT WITH ONE CHILD (VAR. 86)

NFT first grade classrooms. The first quintile corresponded to the 7th lowest score; the second quintile corresponded to the 14th lowest score; the third quintile corresponded to the 21st lowest score; and the fourth quintile corresponded to the 28th lowest score.

Each Follow Through classroom was assigned an implementation score on each implementation variable according to the position of the classroom value among the quintile cutpoints. If the value was greater than the fourth quintile, a score of 5 was assigned; if the value was less than or equal to the fourth quintile, but greater than the third quintile, a score of 4 was assigned. The scores of 3, 2, and 1 were assigned correspondingly. An implementation score of 5 indicates that the Follow Through classroom value exceeded that of at least 80 percent of the NFT classrooms. An implementation score of 4 or 5 indicates that the classroom was on the upper end of the NFT distribution; a score of 3 indicates that the classroom was in the midsection of the NFT distribution; and a score of 1 or 2 indicates that the classroom was on the lower end of the NFT distribution.

When several quintiles had a common value, usually zero, a rule was needed to assign a unique score. The rule was adopted that the highest implementation score possible would be assigned to the classroom. For example, more than 40 percent of the first grade NFT classrooms had a value of zero on the variable Teacher with Two Children. If a Follow Through classroom had a value of zero on this variable, it would fall in the first, second, and third quintiles of the NFT distribution and would be assigned an implementation score of 3.

A total implementation score for a classroom was computed by summing the implementation scores across the corresponding sponsor's implementation variables and then dividing by the highest possible sum. The resulting proportion was then multiplied by 100 so that the total implementation score was expressed in terms of a percentage of the total possible. For example, if a hypothetical sponsor's classroom were being rated on four variables, the highest possible sum of implementation scores for a classroom would be $4 \times 5 = 20$. If a classroom had implementation scores of 3, 3, 4, and 5 on the individual implementation variables, the total implementation score for the classroom would be $(15/20)100 = 75\%$. The reader needs to understand that there is no zero point when computing implementation scores. If all classrooms received the lowest implementation score of 1 on every single implementation variable, their overall implementation score would be 20%, not zero. If they received scores of 5, their overall implementation score would be 100%. Thus, the actual range is from 20 to 100 and the midpoint is a score of 60.

To compare the sponsor programs with Non-Follow Through classrooms, SRI computed a total implementation score for each Non-Follow Through classroom on each sponsor's set of implementation variables. The mean and the standard deviation of the Non-Follow Through pooled classrooms were computed for each sponsor. These statistics serve two purposes:

- The Non-Follow Through mean serves as a reference point for an implementation score. If all the quintile cutpoints had been distinguishable, the mean total implementation score for the Non-Follow Through classrooms would have been 60 (an average score of 3 on each variable out of 5 possible on all variables). Since there are a number of variables for which some or all Non-Follow Through quintiles are zero, it was necessary to compute a separate mean for each set of sponsor-implemented variables.
- The standard deviation was used as a scaling factor to compare a sponsor's total implementation score relative to the Non-Follow Through score. A t-test was used to test whether the mean total implementation score for Follow Through was significantly greater than the mean for Non-Follow Through.

An analysis of variance was run separately for each sponsor and grade level to test whether the sites differed on their mean total implementation scores. This test indicates whether the variability in implementation scores among sites is large relative to the within-site variability among classrooms at a site.

Assessment of the Accuracy of the Implementation Scores

The accuracy of the implementation score assigned to a particular Follow Through classroom for a particular variable depends on the accuracy of the quintile estimates and the accuracy of the estimate of the Follow Through classroom value. Each factor was examined separately because of the intractability of deriving results when examining them simultaneously.

Two factors relate to the stability and accuracy of the quintiles:

- Sampling of class days for each Non-Follow Through classroom.
- Sampling of Non-Follow Through classrooms.

The former is related to the day-to-day variability found in Non-Follow Through classrooms; the latter is related to the number of classrooms

observed and the procedure of sampling classrooms. These factors are relevant when an implementation score is interpreted as an estimate of a Follow Through classroom's position relative to the total population of classrooms that may be considered Non-Follow Through comparisons. Even if we consider the procedure of obtaining implementation scores as a way of scaling the scores of Follow Through classrooms, the stability of such a scale is certainly of interest. We will assume that the Non-Follow Through classrooms are a random sample. This assumption is obviously violated, but it is necessary for the sake of obtaining any notion of the stability of the quincile estimates.

To assess the precision of the quintile estimates, we calculated 95 percent confidence intervals for the quintiles of the estimated classroom distribution, under ideal assumptions concerning the distribution of NFT classrooms. The endpoints of these intervals are displayed in Table 3 for quintile. These computations were based on a sample size of 36, which corresponds to the number of Non-Follow Through classrooms at each grade level. Where the sample size might be reduced because of missing data, the intervals would be slightly longer.

Table 3

95 PERCENT CONFIDENCE INTERVALS FOR THE QUINTILE ESTIMATES EXPRESSED IN PERCENTILES OF THE CLASSROOM DISTRIBUTION

Quintile Cutpoint	Corresponding Percentile	Confidence Interval
1	20	6-31
2	40	22-61
3	60	39-78
4	80	69-94

Consider the first quintile cutpoint as an example of how to interpret the confidence interval. If the experiment of sampling 36 classrooms over 2 or 3 days were replicated many times, then in 95 percent of the replications the first cutpoint (which represents the 20th percentile) will be somewhere between the 6th and the 31st percentile of the Non-Follow Through classroom distribution.

These confidence intervals are rather wide, especially for the second and third quintiles. In terms of the implementation scores, these results indicate that, for a given Follow Through classroom, the true implementation score may be plus or minus one unit from the observed score with a high degree of confidence when we ignore the day-to-day variability of the Follow Through classroom value. That is, if a classroom received a score of 4, there is a good chance that the "true" score might be anywhere between 3 and 5 because of the variability of the quintile estimates.

The effect of the day-to-day variability of the Follow-Through Classroom values on the implementation scores will depend on the position of the classroom value as well as on the magnitude of the day-to-day variability. The stability of the implementation scores was assessed by computing probability distributions of the score for selected values of classroom means and standard errors on selected variables. (The standard error is defined as the standard deviation divided by the square root of the number of days of observation.) Each probability represents the chance that a classroom with a given classroom value and with classroom day-to-day standard deviation will be assigned a given implementation score. The computations were based on two assumptions:

- The deviations from day to day are independent and normally distributed.
- The quintiles are given fixed numbers.

This second assumption has the effect of making the probabilities conditional on the quintile estimates that were derived in the current analysis. The assumption that the estimated classroom value and the quintile estimates are both random is more realistic, but the computations become unwieldy.

Table 4 contains the results of these computations for selected sponsors. The sponsors were selected to provide a range of classroom values. The within-classroom standard deviation for the selected set of CCL and FMO variables were first computed. The estimates for the CCL variables were based on all three days of observation per class; the estimates for the FMO variables were based on the two days of adult-focus observations. The unit of analysis was the day of observation. The estimates were computed separately for each sponsor. Note that the estimates of day-to-day variability are based on three consecutive days of observation in each classroom.

The figures in Table 4 indicate that the implementation scores will generally be within one point of their true value with a high degree of chance. The exceptions will occur when the day-to-day variability is

Table 4

PROBABILITY DISTRIBUTION OF IMPLEMENTATION SCORES
FOR A GIVEN CLASSROOM MEAN AND STANDARD ERROR
(First Grade)

Variables			Class-room	Standard	"True	Probability of Obtaining an Implementation Score of:*				
No.	Name	Sponsor	Value	Error	Score"	1	2	3	4	5
66†	Numbers, math, arithmetic	University of Kansas	21.3	2.0	4	.00	.02	.14	.68	.16
		Bank Street	15.7	5.5	2	.21	.38	.15	.17	.08
		High/Scope	19.8	7.7	4	.13	.23	.11	.21	.32
67‡	Reading, alphabet, language development	University of Arizona	41.8	6.8	2	.33	.42	.22	.02	.01
		University or Oregon	54.6	5.0	4	.00	.05	.43	.31	.21
451a§	Adult academic commands/requests and direct questions to children	Bank Street	7.7	.93	3	.06	.11	.37	.38	.08
		University of Kansas	9.8	1.3	5	.00	.01	.05	.21	.73

*These probabilities represent the chance that a classroom with a given classroom mean and standard error will get any given implementation score. The reliability of the implementation scores may be assessed by examining the chances of attaining the "true score."

†The estimated quintiles for variable 66 are 11.2, 17.0, 19.3, and 23.3.

‡The estimated quintiles for variable 67 are 38.8, 46.4, 54.3, and 58.6.

§The estimated quintiles for variable 451a and 5.2, 6.8, 7.8, and 9.0.

large relative to the range of estimated quintile values and when the classroom value is centrally positioned in the Non-Follow Through distribution. For the three variables we examined, this situation was rare.

In summary, the compound effect of estimating the Follow Through classroom value and estimating the Non-Follow Through quintiles is that the estimated implementation score assigned a particular Follow Through classroom may be different from the true score by as much as two points. In the worst situation, a classroom that has a true implementation score of 3 appears to have a substantial chance of getting an observed score from 1 to 5 on any specific variable. The observation of approximately 4 classrooms per site and 20 classrooms per sponsor in each grade level does mitigate the low reliability of individual classroom scores since we are examining patterns of scores among classrooms. Also, the overall implementation score for each classroom, defined as the sum of the scores across implementation variables, will be much more reliable than the scores for each individual variable.

Illustrations of the Results

The highlights of the results are given here to illustrate how the implementation scores were displayed and evaluated. Some of the results for the University of Kansas and the University of Arizona are included to indicate the contrast in results for different sponsors.

The results for each critical variable were presented for each sponsor as illustrated in Table 5, which shows the results for the University of Kansas on a variable critical to this model: Reading, Alphabet, and and Language Development. The table provides the number of classrooms that received a particular implementation score by site. The number and percent of classrooms over all sites are given for each implementation score in the bottom two lines of the table. First and third grade were reported separately.

In this case, 78 percent of the first grade classrooms and 82 percent of the third grade classrooms had scores of 4 or 5. Only three classrooms at each grade level had scores below 3. Some variability is evident among classrooms within a site, such as the Kansas City, Missouri, first grade classrooms.

Table 6 presents the implementation scores for the University of Arizona program on the variable Child Questions to Adults. This particular variable shows quite a range i degree of implementation among sites in the first grade. None of the first grade classrooms in Des Moines or Newark scored higher than 3, and none of the classrooms in

Table 5

READING, ALPHABET, LANGUAGE DEVELOPMENT (Variable 67)--
UNIVERSITY OF KANSAS

Sites	First Grade Classrooms with Implementation Scores of					Third Grade Classrooms with Implementation Scores of				
	1	2	3	4	5	1	2	3	4	5
NYC P.S. 77X		1			1				1	1
Philadelphia VI, Pa.	1			1	2			1	1	2
Portageville, Mo.					4			1	2	
Kansas City, Mo.		1	1	1	1		1		1	2
Louisville, Ky.					4					4
Total classrooms	1	2	1	2	12		1	2	5	9
Percent of classrooms	6%	11%	6%	11%	67%		6%	12%	29%	53%

Table 6

CHILD QUESTIONS TO ADULTS (Variable 350a)--UNIVERSITY OF ARIZONA

Sites	First Grade Classrooms with Implementation Scores of					Third Grade Classrooms with Implementation Scores of				
	1	2	3	4	5	1	2	3	4	5
Des Moines, Iowa	3	1				2		1	1	
Fort Worth, Texas				1	3			2		2
LaFayette, Georgia				3				1	2	1
Lakewood, N.J.				2	2				3	1
Newark, N.J.		1	3					1	1	2
Lincoln, Nebraska			1		3					4
Total classrooms	3	2	4	6	8	2		5	7	10
Percent of classrooms	13%	9%	17%	26%	35%	8%		21%	29%	42%

three other sites--Fort Worth, LaFayette, and Lakewood--scored lower than four. The high scores across sites in third grade may indicate that this component is more easily implemented in the higher grade levels.

Table 7 presents the summary statistics for the total implementation scores by sponsor and grade level. For all sponsors and grade levels, the average total implementation scores were statistically significantly different from the comparison scores. However, the F tests indicate that there may be differences in level of implementation among sites for several sponsors.

The histograms of total implementation scores for the University of Kansas and NFT classrooms are given in Figure 3. Both the Follow Through and Non-Follow Through distributions are shown to give some impression of the degree of overlap between the two distributions and to show the spread and shape of both distributions.

For the University of Kansas, there is only a slight overlap between the Follow Through and comparison total implementation scores for both grade levels. Table 8 presents the total implementation scores for the University of Kansas classrooms by site and grade level. Although some variation is apparent in the implementation scores among sites at the first grade level, all sites do have total scores that are much higher than those of the comparison classrooms.

Histograms of total implementation scores for the University of Arizona and NFT first grade and third grade classrooms are given in Figure 4. The unshaded histograms are the corresponding distributions for the comparison classrooms. The differences in the position of the Follow Through and comparison classrooms are quite evident. The degree of overlap between the two distributions for both grade levels is also apparent. Table 9 presents the total implementation scores for the University of Arizona classrooms by site and grade level. The differences in scores among sites are quite evident. For the first grade, the site scores range from 55.4 for Newark to 83.3 for Lincoln. Furthermore, all the first grade classrooms at the Newark site had a score below 60. A similar pattern is evident at the third grade level.

Summary

The assessment of implementation is important for policy makers, who must decide whether innovative programs can be effectively disseminated, and for evaluators of program effects, who must verify that the intended treatment was applied. If implementation is defined as the extent that

Table 7

IMPLEMENTATION SCORE ANALYSIS

Sponsor	Grade	Follow Through Model Implementation Score Mean	S.D.	Range High	Range Low	Non-Follow Through* Mean	S.D.	t Test t	t Test p<	F Test F	F Test p<
Far West Laboratory Responsive Educational Program	First (N=20)	78.3	4.4	85.9	71.1	60.3	6.3	11.28	.001	2.65	NS
	Third (N=20)	76.4	7.2	89.6	61.5	59.0	9.4	7.18	.001	3.07	.05
University of Arizona Tucson Early Education Model	First (N=23)	73.6	10.7	89.5	54.0	61.8	7.0	4.99	.001	11.76	.001
	Third (N=24)	72.3	9.1	87.6	52.4	60.7	9.3	4.77	.001	4.75	.01
Bank Street College of Education Approach	First (N=19)	74.8	5.5	82.6	63.0	62.7	6.2	7.12	.001	2.37	NS
	Third (N=19)	69.5	6.0	81.5	60.0	62.4	8.6	3.20	.001	1.71	NS
University of Oregon Engelmann-Becker Model	First (N=19)	78.2	8.1	91.2	71.2	61.0	10.7	6.11	.001	17.61	.001
	Third (N=19)	76.5	9.3	90.6	57.6	60.4	10.5	5.62	.001	.91	NS
University of Kansas Behavior Analysis Approach	First (N=18)	84.6	7.9	96.5	64.7	62.4	8.5	9.22	001	5.14	.01
	Third (N=17)	83.3	6.0	91.8	71.2	61.3	9.3	8.89	.001	2.53	NS
High/Scope Cognitively Oriented Curriculum Model	First (N=18)	76.6	6.0	86.9	66.2	63.7	5.8	7.58	.001	15.59	.001
	Third (N=19)	75.0	6.9	86.2	62.1	63.5	6.8	5.93	.001	27.34	.001
Education Development Center Open Education Follow Through Program	First (N=19)	76.9	11.5	93.0	48.0	61.2	9.6	5.35	.001	7.26	.01
	Third (N=17)	75.4	7.1	85.5	59.1	60.7	10.6	5.18	.001	4.54	.05

*N = 35 classrooms for First Grade
N = 36 classrooms for Third Grade

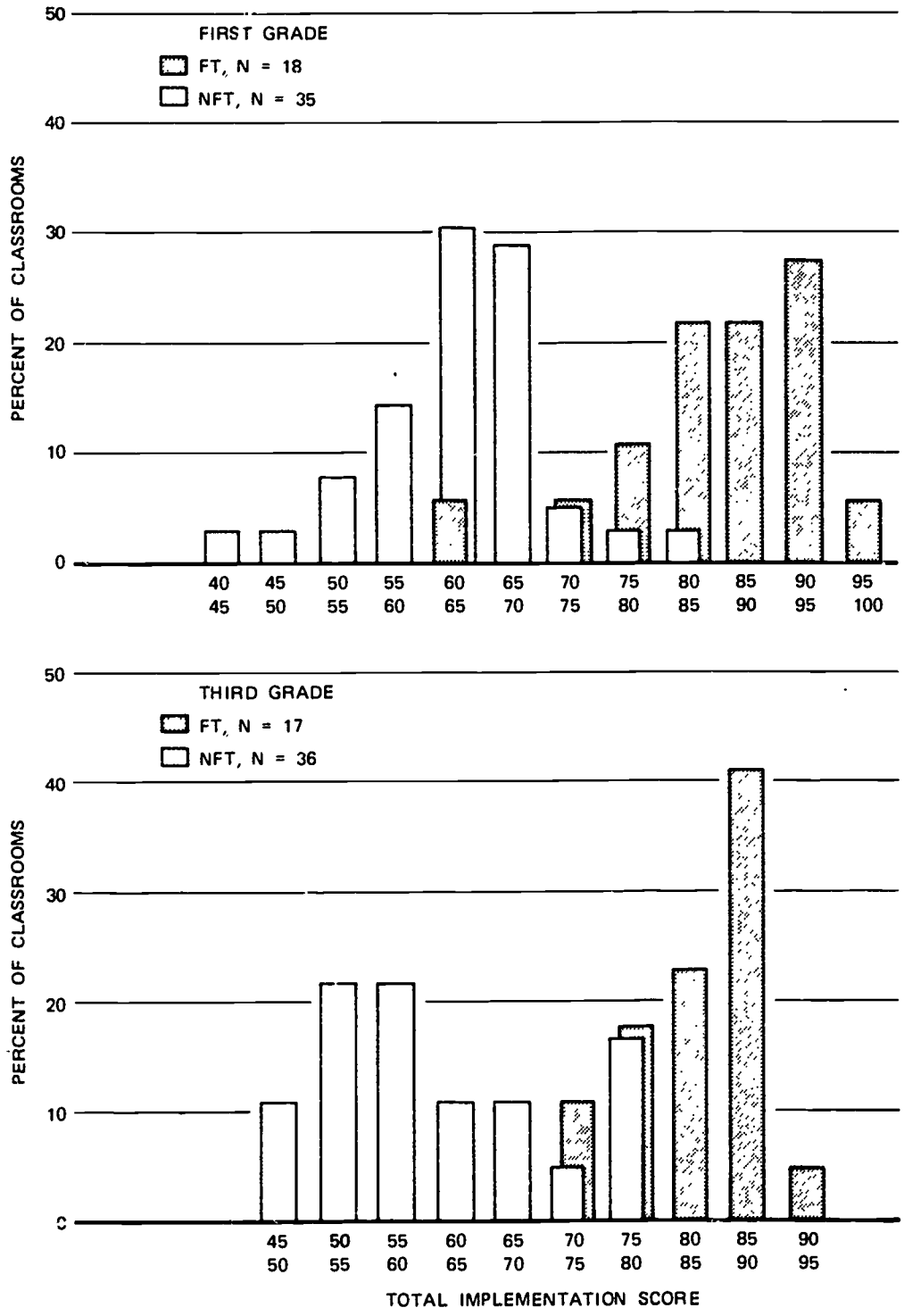

FIGURE 3 HISTOGRAM SHOWING IMPLEMENTATION SCORES FOR UNIVERSITY OF KANSAS

Table 8

TOTAL IMPLEMENTATION SCORES FOR CLASSROOMS BY SITE--UNIVERSITY OF KANSAS

	First Grade					
	Classroom Scores				Site Scores	
Sites	1	2	3	4	$\bar{X}$	S.D.
NYC P.S. 77X (EK)	75.0%	81.3%	%	%	78.1%	4.4
Philadelphia VI (EK)	78.8	90.6	82.4	88.2	85.0	5.4
Portageville (EK)	96.5	91.8	90.6	88.2	91.8	3.5
Kansas City (EK)	82.4	74.1	83.5	64.7	76.2	8.7
Louisville (EK)	85.9	90.6	92.9	85.9	88.8	3.5
Sponsor Scores (N=18):					84.6%	7.9
NFT Scores (N=35):					62.4	8.5
					t = 9.22	
					p < .001	
					f = 5.14	
					p < .01	

	Third Grade					
	Classroom Scores				Site Scores	
Sites	1	2	3	4	$\bar{X}$	S.D.
NYC P.S. 77X (EK)	71.2%	85.0%	%	%	78.1%	9.7
Philadelphia VI (EK)	76.5	82.4	75.3	84.7	79.7	4.5
Portageville (EK)	89.4	74.1	78.8		80.8	7.8
Kansas City (EK)	88.2	88.2	84.7	84.7	86.5	2.0
Louisville (EK)	88.2	91.8	87.1	85.9	88.2	2.5
Sponsor Scores (N=17):					83.3%	6.0
NFT Scores (N=36):					61.3	9.3
					t = 8.89	
					p < .001	
					f = 2.53	
					p < NS	

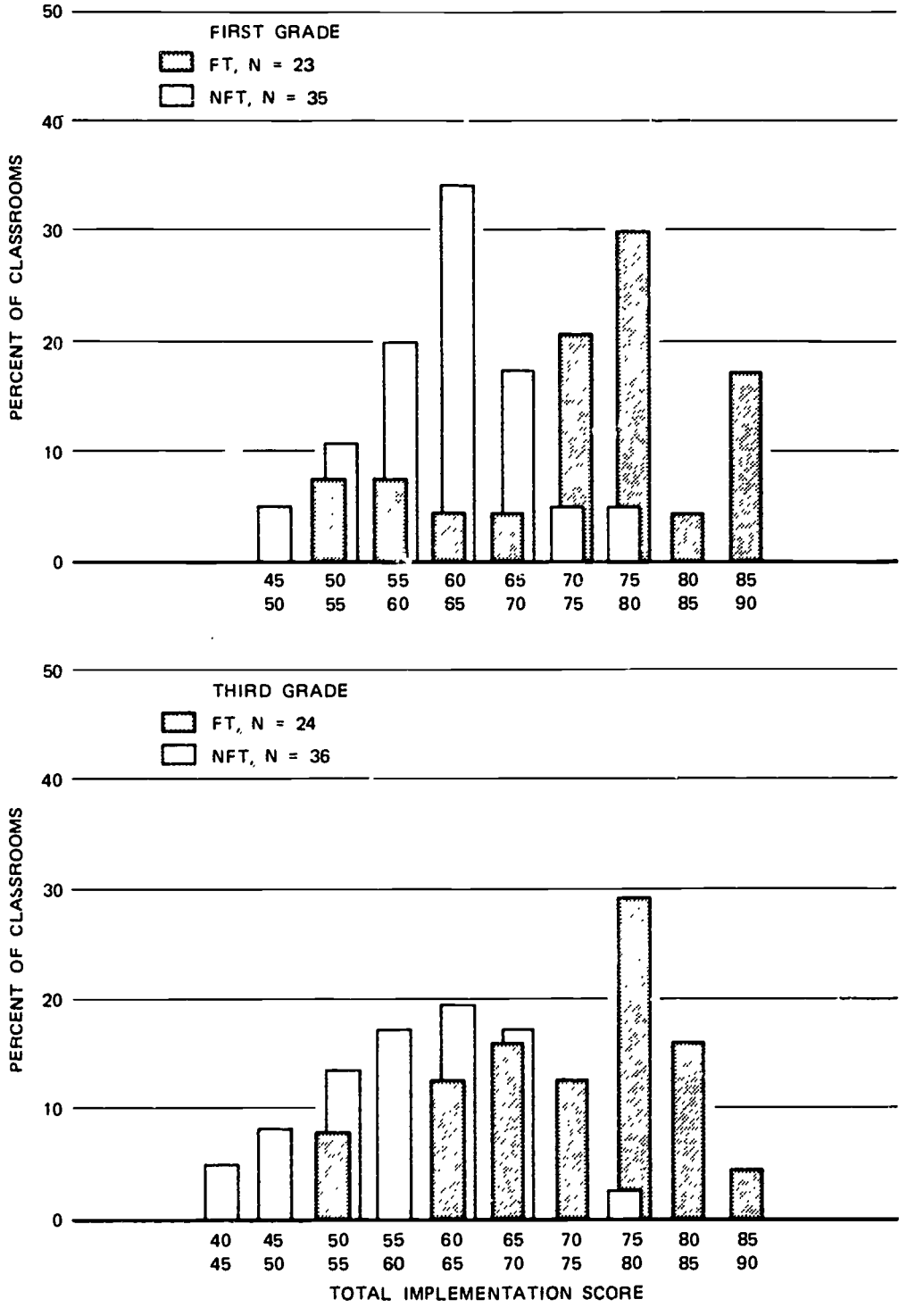

FIGURE 4 HISTOGRAM SHOWING IMPLEMENTATION SCORES FOR UNIVERSITY OF ARIZONA

Table 9

TOTAL IMPLEMENTATION SCORES FOR CLASSROOMS BY SITE--UNIVERSITY OF ARIZONA

First Grade

Sites	Classroom Scores 1	2	3	4	Site Scores $\bar{X}$	S.D.
Des Moines (EK)	79.0%	62.9%	69.5%	71.4%	70.7%	6.7
Fort Worth (El)	85.7	78.1	70.5	75.2	77.4	6.4
LaFayette (El)	79.0	71.4	87.6		79.4	8.1
Lakewood (EK)	78.1	74.3	76.2	79.0	76.9	2.1
Newark (EK)	57.1	54.0	56.2	54.3	55.4	1.5
Lincoln (EK)	89.5	88.6	74.3	81.0	83.3	7.1
Sponsor Scores (N=23):					73.6%	10.7
NFT Scores (N=35):					61.8	7.0
					t = 4.39	
					p < .001	
					f = 11.76	
					p < .001	

Third Grade

Sites	Classroom Scores 1	2	3	4	Site Scores $\bar{X}$	S.D.
Des Moines (EK)	63.7%	52.4%	53.3%	75.2%	61.7%	10.9
Fort Worth (El)	66.7	80.0	82.9	84.8	78.6	8.2
LaFayette (El)	68.6	71.4	73.3	87.6	75.2	8.5
Lakewood (EK)	76.2	78.1	76.2	73.3	76.0	2.0
Newark (EK)	61.9	63.8	67.6	63.8	64.3	2.4
Lincoln (EK)	77.1	75.2	78.1	81.9	78.1	2.8
Sponsor Scores (N=24):					72.3%	9.1
NFT Scores (N=36):					60.7	9.3
					t = 4.77	
					p < .001	
					f = 4.75	
					p < .01	

essential program components are present in the classroom, the steps in an evaluation of implementation include: (1) a determination of the essential program components, (2) a translation of these components in terms of observable phenomena, (3) a measure of the phenomena, and (4) a standard by which to judge implementation.

The SRI approach was formulated in the context of the SRI Classroom Observation Instrument in liaison with the Follow Through sponsors. Only program components that could be translated into observation variables were included in the analysis. To focus on the innovative aspects of the Follow Through programs, the standard established to assess implementation was based on comparisons with Non-Follow Through classrooms.

Implementation scores were derived for individual observation variables, and a total implementation score was derived for each classroom. While all sponsors' classrooms had mean total implementation scores that were significantly different from the Non-Follow Through scores, there were differences in the pattern of implementation among sponsors, as illustrated by the results for the University of Kansas and the University of Arizona.

REFERENCES

Stallings, Jane, and David Kaskowitz, "Follow Through Classroom Observation Evaluation, 1972-1973," SRI Project URU-7370, Stanford Research Institute, Menlo Park, California (August 1974).

Stallings, Jane, "Follow Through Program Classroom Observation Evaluation, 1971-1972," SRI Project URU-7370, Stanford Research Institute, Menlo Park, California (August 1973).

Stearns, Marion S., Kathryn A. Preecs, and Gerald T. Steinmetz, "Classroom Observation Study of Implementation in Head Start Planned Variation, 1970-1971," SRI Project URU-8071, Stanford Research Institute, Menlo Park, California (August 1973).